Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Developing a Composite Measure to Represent Information Flows in Networks: Evidence from a Stock Market

Wuyue (Phoebe) Shangguan, Alvin Chung Man Leung, Ashish Agarwal, Prabhudev Konana, Xi Chen

informs
http://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Articles in Advance, pp. 1–16
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

Developing a Composite Measure to Represent Information Flows in Networks: Evidence from a Stock Market

Wuyue (Phoebe) Shangguan,[a] Alvin Chung Man Leung,[b] Ashish Agarwal,[c] Prabhudev Konana,[d] Xi Chen[e,f,*]

[a] Department of Management Science, School of Management, Xiamen University, Xiamen, Fujian 361005, China; [b] Department of Information Systems, College of Business, City University of Hong Kong, Kowloon, Hong Kong; [c] Department of Information, Risk, and Operations Management, McCombs School of Business, The University of Texas at Austin, Austin, Texas 78712; [d] Robert H. Smith School of Business, University of Maryland, University of Maryland, College Park, Maryland 20742-1815; [e] Department of Data Science and Management Engineering, School of Management, Zhejiang University, Hangzhou 310058, China; [f] Center for Research on Zhejiang Digital Development and Governance, Hangzhou 310058, China
*Corresponding author

Contact: wuyuesg@xmu.edu.cn, https://orcid.org/0000-0002-5150-6955 (W(P)S); acmleung@cityu.edu.hk, https://orcid.org/0000-0001-8961-8357 (ACML); ashish.agarwal@mccombs.utexas.edu, https://orcid.org/0000-0001-9259-0113 (AA); pkonana@umd.edu (PK); chen_xi@zju.edu.cn, https://orcid.org/0000-0001-9776-2914 (XC)

Received: January 30, 2020
Revised: December 17, 2020; July 3, 2021
Accepted: August 12, 2021
Published Online in Articles in Advance: October 12, 2021

https://doi.org/10.1287/isre.2021.1066

Abstract: There is increasing interest in information systems research to model information flows from different sources (e.g., social media, news) associated with a network of assets (e.g., stocks, products) and to study the economic impact of such information flows. This paper employs a design science approach and proposes a new composite metric, eigen attention centrality (EAC), as a proxy for information flows associated with a node that considers both attention to a node and coattention with other nodes in a network. We apply the EAC metric in the context of financial market where nodes are individual stocks and edges are based on coattention relationships among stocks. Composite information from different channels is used to measure attention and coattention. To evaluate the effectiveness of the EAC metric on predicting outcomes, we conduct an in-depth performance evaluation of the EAC metric by (1) using multiple linear and nonlinear prediction methods and (2) comparing EAC with a benchmark model without EAC and models with a set of alternative network metrics. Our analysis shows that EAC significantly outperforms other measures in predicting the direction and magnitude of abnormal returns of stocks. Besides, our EAC specification has better predictive performance than alternative specifications, and EAC outperforms direct attention in predicting abnormal returns. Using the EAC metric, we derive a stock portfolio and develop a trading strategy that provides significant and positive excess returns. Lastly, we find that composite information has significantly better predictive performance than separate information sources, and such superior performance owes to information from social media instead of traditional media.

History: Xiaoquan (Michael) Zhang, Senior Editor; Beibei Li, Associate Editor.
Funding: This work was supported by the National Natural Science Foundation of China [Grants 71931009, 71972168, 72171200, 91546107, and 71821002] and the Research Grants Council of the Hong Kong Special Administrative Region, China [Project CityU 21504216].
Supplemental Material: The online appendices are available at https://doi.org/10.1287/isre.2021.1066.

Keywords: Eigen attention centrality • information flow • attention • coattention • network structure • network analysis • return prediction

1. Introduction

Individuals search, interact, share, or transact on digital platforms, facilitating information flows about an asset or among multiple assets (e.g., products, stocks). In the simplest form, sites such as Amazon.com or Yahoo!Finance use these activities to show users "frequently bought/viewed together," or "customers who viewed this item also viewed," information. In the past, researchers have exploited search-based users' digital footprints on platforms to construct networks as a proxy for coattention among various products or assets[1] and used such network linkages to evaluate market-level outcomes such as stock returns (Agarwal et al. 2017, Leung et al. 2017) and product sales (Lin et al. 2017). However, beyond direct search-based information, there are various sources of information such as news, analyst reports, and social media, which contribute to information flows within the network of linkages, potentially impacting market outcomes. Further, such information can flow directly to assets (i.e., attention) as well as across assets (i.e., coattention). Current network measures such as degree and eigenvector centrality are based only on linkages, but do not represent the extent of information flows. This research adopts a design science perspective to develop a new network measure as a proxy for node-specific and cross-node information flows from various sources and to assess their economic impact within such networks.

There are compelling reasons to model overall information flows from various sources using a network approach. Hirshleifer (2015), in the context of the stock market, argues that social interactions (e.g., virtual communities) and linkages among individuals (e.g., analysts) or assets (e.g., stocks) may provide insights into understanding stock prices since they result in transmission biases and amplification processes that affect outcomes in the stock market. In a product market, such interactions can represent the overall interest and subsequent demand for products. Thus, a composite information flow-based network measure would better represent the complex information environment from different sources and can be used for prediction of prices or demand. To address this, we extend eigenvector centrality (Borgatti 2005) to develop a novel metric, eigen attention centrality (EAC). As a design artifact, EAC can be a proxy for the overall information flows associated with a node in a network and capture varying levels of attention and coattention among nodes. Consistent with the design science principles (Hevner et al. 2004, Gregor and Hevner 2013), we evaluate the effectiveness of EAC by applying it in the context of financial markets to predict abnormal stock returns and comparing the predictive performance with other metrics.

The EAC measure provides several benefits over commonly used network metrics such as degree, closeness, betweenness, and PageRank centralities. First, existing network measures are mainly based on the number of network connections and assume all connections are equal. However, EAC is modeled based on information flows across network linkages where links can channel different levels of information flows. Thus, EAC incorporates such link heterogeneity in intensity of information flows between connections. Second, instead of assuming nodes are identical, EAC assumes that each node in the network has a weight based on the information about that node. Thus, EAC also incorporates the node heterogeneity. Third, EAC allows multiple diffusion paths simultaneously for information flows between two nodes and not just transmission through the shortest path. This is particularly important when the cumulative information flows are important to determine the predictability of outcomes such as asset returns and product demand. Last, the weights associated with a node and between nodes are composite measures of different types of information from various channels including traditional (e.g., news) and new forms of communication channels (e.g., social media). Hence, the EAC metric captures both the attention to a node and information flows across nodes (i.e., coattention) in measuring the relative importance of an individual node in a network.

It is also important to investigate whether the EAC metric can better predict outcomes than other network measures that only consider network connections. Further, studying the predictability of EAC on outcomes (e.g., prediction of returns) allows us to understand the relative impact of different information sources. For example, Luo et al. (2013) show greater impact of social media on stock return predictability as compared with traditional media and attribute it to higher information richness and visibility of social media. Similarly, the predictive performance of information flows based on attention and coattention on asset returns or product demand could also vary depending on media types.

Previous research on information networks has studied the impact of single source of information on many outcomes (e.g., demand, sales) using standard network measures (e.g., degree and PageRank centrality) in contexts such as product recommendation or copurchase networks (Oestreicher-Singer and Sundararajan 2012, Oestreicher-Singer et al. 2013, Dhar et al. 2014, Lin et al. 2017), online collaboration networks (Susarla et al. 2012, Zhang and Wang 2012, Peng and Dey 2013), and labor flow networks (Liu et al. 2020). In the context of stock networks, Leung et al. (2017) and Agarwal et al. (2017) rely on single source of information to construct coattention-based networks. Whereas Drake et al. (2016) consider multiple sources of information, they focus on group-level outcomes such as return comovement and do not evaluate stock return predictability. Further, none of the previous studies compare the importance of information flows from different sources due to attention and coattention across stocks on predicting stock performance (please refer to Online Appendix A for a literature review).

Therefore, to bridge the aforementioned gaps, we address the following research questions: (1) How do we develop a single metric (i.e., EAC) to represent various information flows to nodes in a network? (2) Can the EAC metric enhance predictions of outcomes and how well does it perform relative to other network metrics? (3) What is the differential impact of these information flows from traditional and social media on the predictability of outcomes?

To evaluate the effectiveness of EAC on predictions, we develop an application in a financial context using data from a large financial portal in China. We construct networks of stocks for the entire Chinese stock market[2] over the period from June 2015 to July 2016 and model the stock market as a series of weighted and directed networks based on investors' coattention relationships among stocks. In the networks, nodes are individual stocks and directed edges act as conduits for information flows among connected stocks (i.e., coattention). The weights of nodes represent the amount of direct attention to individual stocks and the weights of edges reflect the intensity of coattention across connected stocks. We use four distinct sources[3] of information from both traditional and social media to represent the composite information flows directly to individual stocks and across different stocks.

We conduct an in-depth predictive analysis to demonstrate the helpfulness of EAC in predicting both the direction and magnitude of abnormal returns in the following manner. First, we use a variety of prediction methods (e.g., linear regressions, support vector regressions, neural networks, decision trees, random forests, and gradient-boosted decision trees) and several evaluation criteria (e.g., mean squared error, mean absolute error, and mean absolute percentage error) to assess the predictive performance on a holdout sample. Second, we compare the predictions based on model settings with and without EAC and find that the predictive accuracy can be significantly improved after including EAC into the models. Then, we compare the predictive accuracy of EAC with alternative network measures (e.g., degree, closeness, betweenness, and PageRank) and consistently find that EAC outperforms these measures under different prediction methods. Moreover, by investigating the confidence intervals for differences in predictive accuracy between various models, we demonstrate that EAC can yield significantly better predictions than other network measures. Finally, based on the results of linear predictive models, we develop a trading strategy for investors where we show that when EAC is relatively low, a strategy that buys stocks with positive sentiment and shorts stocks with negative sentiment will generate significant and positive excess returns.

We also study the importance of different information sources in determining the predictive performance of EAC. The results indicate that EAC based on composite information has better predictive performance than EACs based on separate information sources. Further, we find that EAC based on social media has superior performance as compared with EAC based on traditional media. Finally, we uncover that EACs based on composite and social media information can significantly outperform corresponding attention measures in predicting abnormal returns. This suggests that our EAC metric, which considers both attention and coattention, can significantly improve the model's predictive power on abnormal returns.

Our study makes several contributions. First, we add to the emerging information system (IS) literature on the analysis of information flows through digital networks of products and assets. In particular, we develop a new composite metric, that is, EAC, which is different from existing network-based measures that focus solely on connections and ignore the information content in networks. Aligned with design science principles, we design EAC as a proxy for information flows associated with attention and coattention in a network and conduct an in-depth performance evaluation of EAC. This measure has greater predictive power over existing measures. Second, we contribute to research on information diffusion in IS and finance (e.g., Luo et al. 2013, Chen et al. 2014) that focuses on the use of social media to determine stock performance. We establish that coattention through social media can provide greater predictive gains over direct attention, thus extending this stream of research by demonstrating the role of both attention and coattention through social media in predicting abnormal returns. For practitioners, we provide an approach to model information flows for assets and use it to predict demand or returns. In the stock market context, our study can guide investment managers to model information flows using various sources of information and focus on intensity of attention and coattention to devise an appropriate trading strategy.

2. Research Design

Central to our research is EAC, which measures the importance of a node in a network with different levels of direct attention and coattention across nodes. We begin by introducing the concept of EAC and then describe an example from the stock market to show the construction of EAC.

2.1. Concept of EAC

Consider a weighted directed network represented by $G = \{N, E, W_N, W_E\}$, where N is a set of nodes, E is a set of directed edges reflecting coattention between nodes, W_N is a set of node weights capturing the node-specific attention intensity, and W_E is a set of edge weights reflecting the intensity of coattention.[4] Our EAC metric is the eigenvector centrality of the defined doubly weighted network. The key to calculating EAC is the attention intensity (i.e., node weight) and coattention intensity (i.e., edge weight).

2.1.1. Attention Intensity. Users pay varying level of attention to different assets through different channels. As users can acquire information from multiple channels, a *composite attention* index is needed to represent the attention intensity.

2.1.2. Coattention Intensity. Due to limited attention and heterogeneous preferences, users tend to pay attention to a limited set of assets at the same time (i.e., coattention). Information portals and e-commerce platforms can capture this information by monitoring the coattention frequency across assets of individual users.

2.1.3. EAC: Eigenvector Centrality of Doubly Weighted Networks. Based on the abovementioned definitions, let $\mathbf{x}^k = [x_1^k, \ldots, x_n^k]'$ be the vector of eigenvector centrality of assets in network G^k (k indexes network) during a fixed time interval.[5] By solving the largest eigenvalue of the adjacency matrix for network G^k, the corresponding eigenvector $\mathbf{x}^k$ is a vector with nonnegative elements (Newman 2010). Thus, the i th element of $\mathbf{x}^k$ (i.e., x_i^k) is the EAC metric of asset i in network G^k. More details about the mathematical modeling to compute EAC are displayed in Online Appendix B.

2.2. EAC for Stocks

Here we describe different elements of the EAC measurement for individual stocks in the stock market. Investors pay attention to one or more stocks (i.e., coattention) simultaneously. Further, coattention to stocks can be observed on financial web portals such as Yahoo!Finance and Sina.com. This coattention can be used to represent a set of directed edges connecting stocks (i.e., an edge from stock j to stock i indicates when individuals pay attention to stock j, a substantial number of them also pay attention to stock i).

2.2.1. Node Weight: Composite Attention Intensity to Stocks. In behavioral finance, attention is described as a limited resource that reflects the extent to which investors assimilate information of a financial asset (Hirshleifer and Teoh 2003). In the past, traditional media such as news (Barber and Odean 2008) and analyst reports (Drake et al. 2016) have been used to measure the level of attention. Recently, however, there is increasing reliance on new channels such as stock message boards and social networks to access and share information (Antweiler and Frank 2004, Sabherwal et al. 2011, Hirshleifer 2015, Zhang and Zhang 2015). To account for information acquisition from different channels, we derive a composite attention index based on the following proxies of attention: (1) the number of reads of stock-related posts, (2) the number of investor comments on stock-related posts, (3) the number of related news articles, and (4) the number of related analyst reports.[6] Following Drake et al. (2016), we conduct a factor analysis of the four attention proxies and retain the first principal factor[7] that captures the largest variation as the composite attention index. Suppose there are multiple networks constructed in different time periods with each time period having a fixed time interval, then for each network G^k (indexed by k), we compute the value-weighted average[8] of the daily composite attention for stock i in network G^k as $Attention_i^k$ and assign the min-max-normalized value of $Attention_i^k$ as the node weight.

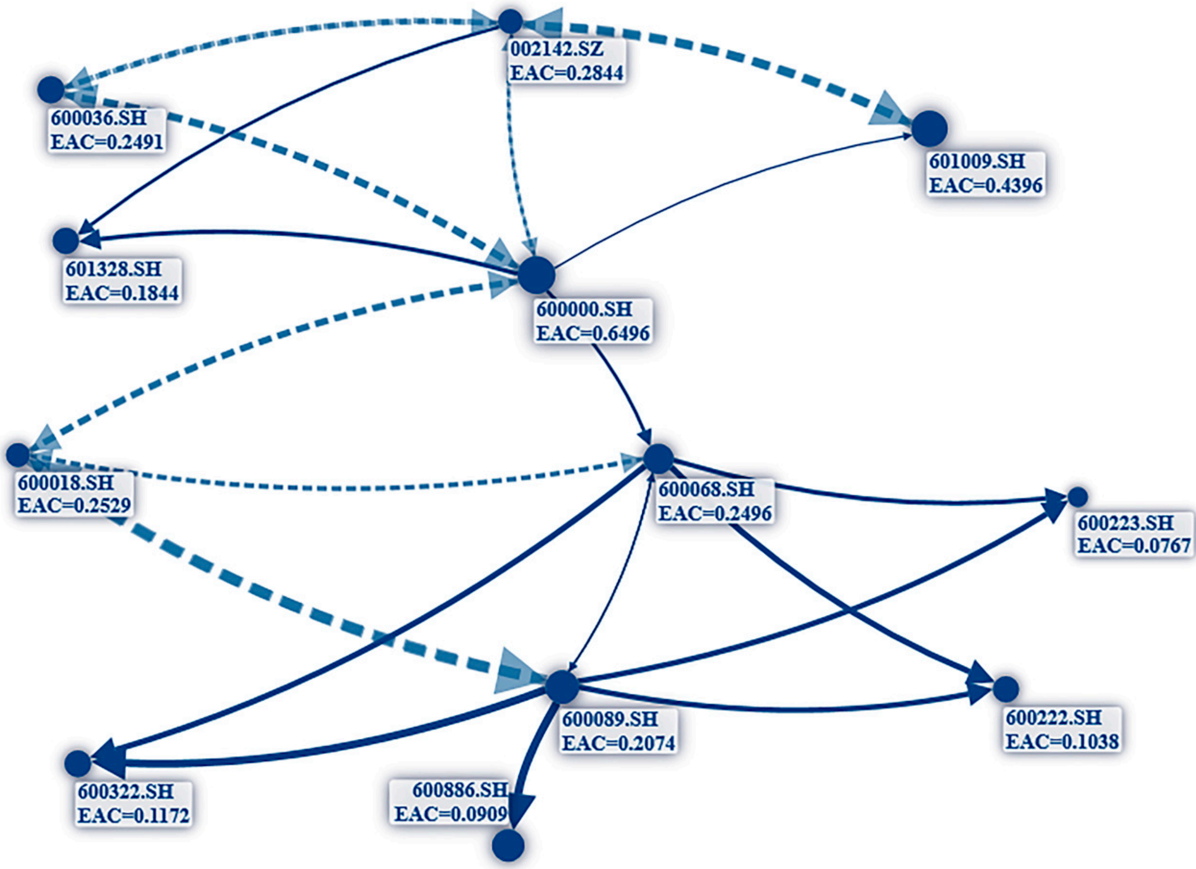
2.2.2. Edge Weight: Coattention Intensity Between Stocks. Previous studies have used cosearch of stocks to reveal the coattention relationships and determine what stocks are cosearched together and the direction of coattention (e.g., Agarwal et al. 2017, Leung et al. 2017). However, the intensity of coattention between stocks is not revealed. It is possible that investors have higher coattention intensity on some stocks than others. Hence, these stocks may experience higher extent of information diffusion, which influences the predictability of stock prices or returns. In our setup, the information portals do not reveal the coattention frequency but only whether two stocks receive coattention. However, once two assets receive attention from overlapped groups, their attention is also likely to be correlated. Therefore, the correlation between attention to individual assets can be used as a proxy for coattention intensity.

We infer the level or intensity of coattention between stocks by adopting an approach similar to Drake et al. (2016). Specifically, for each directed edge $j \rightarrow i$ representing a coattention link from stock j to i in network G^k that is formed during a fixed time interval, we run the following univariate regression using ordinary least squares (OLS):

$$Attention_{i,t} = \alpha_{\mathrm{i}} + \beta_{ij} Attention_{j,t} + \varepsilon_{i,t} \tag{1}$$

where $Attention_{i,t}$ is the daily composite attention index for stock i. We calculate the coefficient of determination, the R-squared (i.e., R_{ij}^2) of Equation (1), which reflects the extent to which attention to stock i can be explained by the attention paid to stock j during the network construction period. We assign the value of R_{ij}^2 as the edge weight for $j \rightarrow i$, which is a proxy of coattention intensity between stocks j and i. Note that the intensity of coattention is only determined for stock pairs where we know about the existence of coattention.[9] We have further verified that the R-squared is significantly lower when there is no coattention on our dataset obtained from Sina Finance. Additionally, this measure can be easily replaced by coattention frequency when available. Figure 1 shows an example of a reduced coattention network of 12 stocks and their EAC values. We can find that the node "600000.SH" has the largest EAC value and occupies the most "central" position in this network. It plays an important role as a structural hole of the network: without this stock, information cannot flow between the top and the bottom halves of the network. Also, stocks with in-links from "600000.SH" tend to have relatively higher EAC. This result is consistent

Figure 1. (Color online) EAC Indices in a Coattention Network



Notes. The node size is proportional to the node weight, which reflects the intensity of attention associated with a stock, and the edge weight is revealed by the thickness of the edge, illustrating the intensity of coattention across connected stocks. Edges marked as dashed are bidirectional.

with the argument of eigenvector centrality that an individual node in a network will be more important if its connections are also from important ones.

3. Data and Coattention Networks

We collected coattention-based stock data from Sina Finance, which owns the biggest market share among all the Chinese financial portals. Based on users' coattention relationships to stocks, Sina Finance determines the coattention list by identifying another nine stocks that have the highest coattention frequencies with the focal stock.[10] Further, Sina Finance identifies the coattention lists solely based on coattention frequencies.[11] Figure 2 shows a screenshot of a sample coattention list.

We collected daily coattention data of all the A-share stocks listed in both Shanghai and Shenzhen Stock Exchanges at 4 p.m. Beijing time every day during the period from June 1, 2015, to July 8, 2016. We obtained daily news coverage, analyst reports, data of individual stock characteristics (e.g., stock returns, market capitalization, trading volume, turnover, and

Figure 2. (Color online) A Screenshot of Coattention List for "600000.SH"

price-to-book ratio), industry returns, and Fama-French risk factors from the China Stock Market & Accounting Research (CSMAR) database. With the help of Chinese Research Data Services (CNRDS) platform, we collected daily data on the number of reads and the number of comments from a popular online stock forum, eastmoney.com.[12] The CNRDS platform also labels the sentiment of each forum post related to a focal stock as positive, negative, or neutral.

We used coattention lists to construct stock networks. A sliding window approach was adopted to capture the dynamic changes in network structure. The window size was fixed as four weeks and the sliding size was set as one week. Specifically, we began to use the initial four-week coattention data to construct the first network. After that, we moved the window one week forward and constructed the second network using the updated data. This procedure was repeated until the most recent network was constructed. As a result, we had a total of 54 coattention networks from June 1, 2015, to July 3, 2016. We also conducted a simple statistical analysis of these networks and found that the networks were relatively stable in terms of the number of nodes. A summary of network characteristics is provided in Online Appendix C.

4. Methodology

4.1. Prediction Models

To demonstrate the effectiveness of EAC, we evaluate whether EAC can be used to predict abnormal stock returns. Abnormal returns represent firm equity value beyond what is expected based on a set of market risk factors using asset pricing models (please refer to Online Appendix D.1 for calculation details). Abnormal returns are commonly used in the finance and IS literature (e.g., Dewan and Ren 2007, Luo et al. 2013, Chen et al. 2014) to predict stock performance. Note that the abnormal returns can be either positive or negative, indicating the direction in which the actual returns deviate from the market expectation. Thus, we also employ the absolute values of abnormal returns to study the predictability of EAC on the magnitude of abnormal returns.

Table 1 presents the description of the variables. In addition to the main predictor EAC, we include other known determinants of abnormal returns (e.g., raw returns, industry returns, market capitalization, trading volume, and price-to-book ratio). As investor sentiment may influence stock performance as well (Antweiler and Frank 2004, Brown and Cliff 2004, Baker and Wurgler 2007, Sabherwal et al. 2011), we calculate and account for the sentiment index proposed by Antweiler and Frank (2004) (please refer to Online Appendix D.2 for calculation details). Further, because stocks within the same habitat exhibit strong return comovement (Wahal and Yavuz 2013), we also compute the value-weighted abnormal returns of the peer stocks to control for return predictability due to comovement. The descriptive statistics and the correlation matrix for the variables are shown in Online Appendix E.

First, we develop the following baseline model to study the predictability of EAC on abnormal returns. We fit the linear model using Fama-MacBeth two-step regressions (Fama and MacBeth 1973) with Newey-West adjusted standard errors (Newey and West 1987) to account for lag-one autocorrelation. This procedure is widely adopted in studying return predictability (e.g., Pontiff and Woodgate 2008, Menzly and Ozbas 2010, Belo et al. 2014), which considers both interfirm dependencies and intrafirm serial correlations:

$$\begin{aligned} AbnReturn_{it} = {} & \beta_0 + \beta_1 EAC_{i,t-4:t-1} + \beta_2 AbnRetPeers_{i,t-4:t-1} \\ & + \beta_3 Sentiment_{i,t-4:t-1} + \beta_4 \overline{Return}_{i,t-4:t-1} \\ & + \beta_5 \overline{IndReturn}_{i,t-4:t-1} + \beta_6 \overline{MktCap}_{i,t-4:t-1} \\ & + \beta_7 \overline{Volume}_{i,t-4:t-1} + \beta_8 \overline{Turnover}_{i,t-4:t-1} \\ & + \beta_9 \overline{P2B}_{i,t-4:t-1} + \varepsilon_{i,t} \end{aligned} \tag{2}$$

Although linear models have good interpretability of results, nonlinear models tend to yield better predictive outcomes especially when nonlinearity exists among predictors. Hence, we also use EAC to predict abnormal returns by applying the following nonlinear methods: (1) support vector regression (SVR) with polynomial kernels; (2) neural networks based on multilayer perceptron (MLP); (3) decision trees (DTs) and ensemble methods including (4) random forests (RFs) and (5) gradient-boosted decision trees (GBDTs) for regressions.

Since nonlinear methods tend to have a higher level of model complexity which increases the risk of overfitting, we employ various approaches to eliminate overfitting. First, we introduce the regularization terms to penalize the training loss. Second, when using tree-based algorithms, we control for the scale (e.g., minimum number of samples per leaf and per split) and the maximum depth of the trees to avoid learning too fine rules that may lead to overfitting. Third, in addition to individual base models, we also use ensemble methods such as random forests and GBDTs that are more robust to overfitting (Dietterich 2000, Rokach 2010, Kuncheva 2014). Last, we also compare the values of training errors and testing errors for different models and show that they are relatively close, so it is less likely to have overfitting issue (please refer to Online Appendix F.1).

4.2. Model Evaluation Strategy

Our validation strategy is to test the predictive accuracy in holdout samples. Following the same approach as in Section 3, we used the coattention data in the year 2017 from Sina Finance and formed 49 weighted

Table 1. Variable Definitions

Variables	Definitions
Outcomes	
$AbnReturn_{it}$	Abnormal returns for stock i on week t
$\lvert AbnReturn_{it}\rvert$	Absolute values of abnormal returns for stock i on week t
Predictors	
$EAC_{i,t-4:t-1}$	The composite eigen attention centrality for stock i in the past four weeks
$Attention_{i,t-4:t-1}$	The composite index of investor attention for stock i in the past four weeks
$AbnRetPeers_{i,t-4:t-1}$	Value-weighted average of abnormal returns in the past four weeks for stocks connecting with focal stock i
$\overline{Sentiment}_{i,t-4:t-1}$	Investor sentiment index for stock i in the past four weeks
$\overline{Return}_{i,t-4:t-1}$	Average raw returns for stock i in the past four weeks
$\overline{IndReturn}_{i,t-4:t-1}$	Average industry returns for stock i in the past four weeks
$\overline{MktCap}_{i,t-4:t-1}$	Average market capitalization for stock i in the past four weeks
$\overline{Volume}_{i,t-4:t-1}$	Average trading volume for stock i in the past four weeks
$\overline{Turnover}_{i,t-4:t-1}$	Average turnover rate for stock i in the past four weeks
$\overline{P2B}_{i,t-4:t-1}$	Average price-to-book ratio for stock i in the past four weeks

networks. Then we recalculated the EACs and other variables as defined in Table 1. We use various evaluation criteria such as root mean squared error (RMSE) or mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) to test the out-of-sample predictive accuracy. Specifically, we assess the predictive performance of EAC by (1) comparing model settings with and without EAC, (2) comparing EAC with existing centrality measures such as degree, closeness, betweenness, and PageRank centralities, (3) comparing our EAC specification with alternative specifications, (4) showing the robustness of our findings using different matrix normalization methods and censored data, and (5) developing a trading strategy based on the predictive relationship between EAC and abnormal returns.

5. Results

5.1. Performance of EAC

We first investigate if the EAC metric can provide additional predictive gains compared with benchmark models without EAC. Then we extend this benchmark for model comparisons by comparing the EAC metric with alternative network measures such as degree, closeness, betweenness, and PageRank centralities and comparing EAC with alternative specifications. Table 2 shows the evaluation results for different models. Overall, we find that nonlinear methods perform better than the baseline linear model, and the results are consistent across both linear (panel A) and nonlinear models (panels B–F) that indicate the EAC metric can yield better predictions. Specifically, models with EAC have much lower RMSE, MAE, and MAPE on the holdout sample compared with the benchmark models without EAC. This suggests that incorporating EAC into the model can greatly improve the predictive accuracy of abnormal returns. In addition, models with EAC also have smaller forecasting errors than models with alternative centrality measures. This suggests that EAC also outperforms other centrality measures in predicting abnormal returns. Finally, our EAC specification performs better than alternative specifications such as unweighted and/or ego-centric networks (please refer to Online Appendix F.2) and the results remain consistent under a different matrix normalization method (please refer to Online Appendix F.3).

5.2. Differences in Predictive Accuracy

To reflect the extent to which EAC outperforms other measures, we investigate if the improvement in predictive accuracy due to EAC is economically significant. Previous research advises on examining whether the size of differences in predictive accuracy is notable between various models because even a small visual improvement in prediction may cause considerable economic impact (Dhar et al. 2014). Therefore, following Dhar et al. (2014), we compute the 95% confidence intervals for differences in predictive accuracy between the EAC model and each of the alternative models. We adopt a bootstrapping procedure by repeatedly drawing 1,000 random samples to calculate the confidence intervals for the differences in MSE and MAE, respectively.[13] Table 3 shows the results.

From Table 3, we can find that the lower limits of all the confidence intervals for the differences between an alternative model and the EAC model are greater than zero. This suggests that we can be 95% confident that the model with EAC yields significantly better predictions than the alternative models. Furthermore, confidence intervals also provide information on how large the difference is. For example, based on the results of predicting the original values of abnormal returns using SVR in panel A, we are 95% confident that the difference in MSE between the benchmark model without EAC and the EAC model falls within the range of 1.4564 to 1.7906. Besides, we find that these intervals are relatively narrow, which suggests that

Table 2. Predictive Accuracy for Different Models

Panel A. Fama-MacBeth regressions with Newey-West adjusted standard errors

Models	Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it}\rvert$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
EAC	7.9660	7.2500	25.9380	5.4970	4.3320	13.7490
Without EAC	12.6230	12.0204	45.1440	8.6425	7.7669	28.4250
Degree	18.7011	18.2711	68.7022	9.8840	9.0660	33.3540
Closeness	10.7450	10.1400	37.2210	5.5400	4.3790	13.9170
Betweenness	9.3100	8.6540	31.5690	5.8480	4.7230	15.5990
PageRank	10.1720	9.5520	35.0080	5.7030	4.5590	14.8960

Panel B. Support vector regressions with polynomial kernels

Models	Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it}\rvert$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
EAC	3.9900	2.7292	1.5934	3.2680	1.9517	3.8105
Without EAC	4.1618	3.0254	5.2165	4.9742	4.0411	6.1005
Degree	4.0735	2.8923	3.9296	4.7700	3.7875	5.1701
Closeness	4.0097	2.7815	2.5822	3.7493	2.4468	5.9251
Betweenness	4.0169	2.7961	2.7750	3.4554	2.1128	5.8917
PageRank	4.0162	2.7948	2.7687	4.5542	3.5140	4.1367

Panel C. Neural networks using multilayer perceptron

Models	Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it}\rvert$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
EAC	4.1790	2.9497	3.9612	3.0642	1.8595	6.0046
Without EAC	4.3880	3.0520	4.9403	5.0044	4.0751	7.4483
Degree	4.8423	3.5114	6.6685	4.8898	3.9659	6.9037
Closeness	4.6497	3.2981	5.0431	3.6005	2.1942	8.0221
Betweenness	4.9123	3.6193	7.2488	4.7283	3.7646	6.3776
PageRank	4.7915	3.4561	6.2298	3.6551	4.0146	6.6676

Panel D. Decision trees

Models	Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it}\rvert$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
EAC	3.9994	2.7387	1.5215	3.1172	1.9898	6.2727
Without EAC	4.2906	3.1268	6.1747	3.3301	2.7411	14.8684
Degree	4.1153	2.9300	3.9952	3.3351	2.7587	14.8254
Closeness	4.0250	2.7996	2.5568	3.2627	2.6271	13.8979
Betweenness	4.0818	2.8794	3.4566	3.9828	2.5526	11.2794
PageRank	4.7170	2.9538	4.3793	3.1678	2.5155	12.9926

Panel E. Random forests

Models	Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it}\rvert$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
EAC	4.0028	2.7340	1.3872	2.8849	2.0679	9.7918
Without EAC	4.3556	3.2505	6.8491	3.0038	2.2942	11.3968
Degree	4.3096	3.1781	5.3645	2.9755	2.2465	10.9685
Closeness	4.1171	2.9266	3.9180	2.9654	2.2286	11.4375
Betweenness	4.1967	3.0276	4.8351	2.9674	2.2315	11.4378
PageRank	4.1560	2.9795	4.3117	2.9387	2.1715	10.5432

Panel F. Gradient-boosted decision trees

Models	Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it}\rvert$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
EAC	4.0023	2.7314	1.0332	2.9219	2.1438	10.6643
Without EAC	4.3383	3.1364	5.8726	3.1036	2.4530	13.0837
Degree	4.1162	2.9146	3.9611	3.0997	2.4480	13.0430
Closeness	4.1468	2.9470	4.4798	3.0884	2.4332	12.9159
Betweenness	4.1374	2.9356	4.3259	3.0853	2.4291	12.8905
PageRank	4.1780	2.9821	4.9243	3.1224	2.4836	13.2154

Table 3. Bootstrapping Confidence Intervals for Mean Differences in Predictive Accuracy

Panel A. EAC vs. without EAC

	(1) Predicting $AbnReturn_{it}$				(2) Predicting $\lvert AbnReturn_{it} \rvert$			
Method	MSE(without EAC) − MSE(EAC)		MAE(without EAC) − MAE(EAC)		MSE(without EAC) − MSE(EAC)		MAE(without EAC) − MAE(EAC)	
Fama-MacBeth	37.0608	37.8852	2.3199	2.3693	10.6633	11.3123	0.8218	0.8623
SVR	1.4564	1.7906	0.2893	0.3248	12.6396	12.9534	2.0663	2.0988
MLP	1.0910	1.4332	0.0792	0.1187	13.5329	13.8459	2.1646	2.1953
DT	2.0079	2.3410	0.3523	0.3895	2.2422	2.3982	0.7859	0.8086
RF	2.9568	3.2764	0.4976	0.5331	0.8428	0.9663	0.2266	0.2501
GBDT	2.5953	2.9333	0.3830	0.4200	1.2196	1.3434	0.3116	0.3320

Panel B. EAC vs. degree

	(1) Predicting $AbnReturn_{it}$				(2) Predicting $\lvert AbnReturn_{it} \rvert$			
Method	MSE(degree) − MSE(EAC)		MAE(degree) − MAE(EAC)		MSE(degree) − MSE(EAC)		MAE(degree) − MAE(EAC)	
Fama-MacBeth	296.6301	298.2671	11.8506	11.9069	93.5223	94.3694	5.3059	5.3440
SVR	0.6816	1.0012	0.1516	0.1897	10.7954	11.1161	1.8121	1.8458
MLP	4.8225	5.1872	0.5131	0.5559	12.4528	12.7413	2.0541	2.0866
DT	0.9274	1.2660	0.1803	0.2168	2.4429	2.6095	0.8101	0.8359
RF	2.5531	2.8718	0.4230	0.4604	0.6193	0.7408	0.1780	0.1998
GBDT	0.7756	1.1169	0.1629	0.1986	1.1898	1.3211	0.3049	0.3271

Panel C. EAC vs. closeness

	(1) Predicting $AbnReturn_{it}$				(2) Predicting $\lvert AbnReturn_{it} \rvert$			
Method	MSE(closeness) − MSE(EAC)		MAE(closeness) − MAE(EAC)		MSE(closeness) − MSE(EAC)		MAE(closeness) − MAE(EAC)	
Fama-MacBeth	42.8003	43.6772	2.6332	2.6865	0.6238	1.1821	0.0559	0.0967
SVR	0.0744	0.3807	0.0346	0.0729	2.8019	3.0719	0.4728	0.5047
MLP	3.2296	3.5748	0.3036	0.3431	4.5291	4.7153	0.3252	0.3542
DT	0.1435	0.4694	0.0450	0.0846	1.6484	1.8088	0.6560	0.6805
RF	0.8105	1.1063	0.1638	0.2014	0.4588	0.5758	0.1540	0.1763
GBDT	1.0171	1.3346	0.1941	0.2306	1.1162	1.2395	0.2899	0.3118

Panel D. EAC vs. betweenness

	(1) Predicting $AbnReturn_{it}$				(2) Predicting $\lvert AbnReturn_{it} \rvert$			
Method	MSE(betweenness) − MSE(EAC)		MAE(betweenness) − MAE(EAC)		MSE(betweenness) − MSE(EAC)		MAE(betweenness) − MAE(EAC)	
Fama-MacBeth	24.3077	25.0842	1.5906	1.6433	6.1704	6.7641	0.4953	0.5330
SVR	0.1302	0.4397	0.0526	0.0885	0.9429	1.1854	0.1408	0.1739
MLP	5.6000	5.9934	0.6283	0.6666	11.0740	11.3584	1.8555	1.8883
DT	0.6453	0.9553	0.1249	0.1651	4.0250	4.3058	0.4523	0.4862
RF	1.4968	1.7922	0.2682	0.3028	0.4784	0.6044	0.1583	0.1784
GBDT	0.9541	1.2661	0.1812	0.2199	1.1023	1.2276	0.2862	0.3072

Panel E. EAC vs. PageRank

	(1) Predicting $AbnReturn_{it}$				(2) Predicting $\lvert AbnReturn_{it} \rvert$			
Method	MSE(PageRank) − MSE(EAC)		MAE(PageRank) − MAE(EAC)		MSE(PageRank) − MSE(EAC)		MAE(PageRank) − MAE(EAC)	
Fama-MacBeth	34.3435	35.2147	2.1694	2.2218	3.4599	4.0971	0.2832	0.3240
SVR	0.1297	0.4393	0.0508	0.0865	8.9185	9.2099	1.5372	1.5729
MLP	4.0486	4.4032	0.4391	0.4791	2.7050	2.9253	2.1058	2.1398
DT	0.3230	0.6238	0.0616	0.1022	1.2181	1.3738	0.5580	0.5857
RF	1.1722	1.4877	0.2169	0.2590	0.3413	0.4760	0.0988	0.1201
GBDT	1.2402	1.5790	0.2277	0.2675	1.3755	1.5007	0.3437	0.3661

the estimate of the mean differences in predictive accuracy is relatively precise.

5.3. Sensitivity to Censored Data

Our analysis could be affected by two separate censoring issues. We assume the existence of coattention based on the information available from Sina Finance, which restricts the coattention information to the top nine stocks. However, there may be other coattention relationships not revealed by the platform. Similarly, the directionality of coattention can be censored as a stock may have another stock in its top nine coattention list but the reciprocal coattention is outside the top nine list. To show that the EAC metric is robust and insensitive to censored data, we reconstruct the networks based on (1) only bidirectional edges and (2) the top five stocks in the coattention lists. We recalculate the EAC metric in each of the networks and assessed the predictive accuracy of EAC on abnormal returns in the holdout set. Based on Table F6 and Table F7 of Online Appendix F.5, we find that under either situation, models with EAC still yield better predictions compared with other models.

5.4. Trading Strategy

To further assess the helpfulness of EAC in return predictions, we develop a trading strategy using EAC under which the investment portfolio can help investors earn excess gains. To achieve this goal, we first estimate the models in Equation (2) and employ the linear model results to develop the trading strategy. Table 4 shows the estimation results.

From Table 4, we find that the coefficient of $EAC_{i,t-4:t-1}$ on $AbnReturn_{it}$ in model (1) is −5.257 and significant at 0.01 level. In model (2), the coefficient of $EAC_{i,t-4:t-1}$ on $|AbnReturn_{it}|$ is also negative and significant, which shows that a 1% increase of EAC is associated with a 4.44% decrease in the magnitude of abnormal returns. The results indicate that a higher level of EAC leads to a smaller magnitude of future abnormal returns. This suggests that a stock with increased EAC may be experiencing faster information diffusion across stocks. This results in an increased level of price efficiency, as reflected by a lower magnitude of future abnormal returns.

In addition, we find that stocks with positive sentiment can yield significantly higher future abnormal returns. Thus, we can construct an investment portfolio that buys stocks with positive sentiment and shorts stocks with negative sentiment.[14] Another objective is to employ EAC in this trading strategy. Because low EAC is associated with slow price discovery, we expect that low-EAC scenarios can benefit arbitragers from the slow change of stock prices. Hence, we anticipate that under the situation of a lower level of EAC, the portfolio based on sentiment can earn significantly higher excess returns relative to the high-EAC scenarios.

We use the holdout sample in year 2017 to assess the trading strategy.[15] Table 5 reports the gains of the trading strategy by double sorting all stocks based on EAC and sentiment. In panel A (panel B), stocks are sorted according to the past-one-month EAC into terciles (quintiles).[16] Stocks within each tercile (quintile) are further split into three portfolios based on their last-month investor sentiment: positive, negative, and neutral sentiment stocks. We then calculate the value-weighted average of weekly returns for all the portfolios. Under each EAC tercile (quintile), we construct an investment portfolio that buys positive sentiment stocks and shorts negative sentiment stocks and calculate the returns of the buy-short portfolio as the difference between positive and negative sentiment portfolios. Portfolios are held for one week and

Table 4. Estimation Results for Fama-MacBeth Regressions

Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it} \rvert$		
Variables	Coef.	Std. error	Variables	Coef.	Std. error
$EAC_{i,t-4:t-1}$	−5.257**	(1.680)	$EAC_{i,t-4:t-1}$	−4.443**	(1.257)
$AbnRetPeers_{i,t-4:t-1}$	0.894***	(0.163)	$\lvert AbnRetPeers_{i,t-4:t-1}\rvert$	0.481**	(0.139)
$\overline{Sentiment}_{i,t-4:t-1}$	0.412***	(0.080)	$\lvert\overline{Sentiment}_{i,t-4:t-1}\rvert$	0.071	(0.039)
$\overline{Return}_{i,t-4:t-1}$	−0.230***	(0.040)	$\lvert\overline{Return}_{i,t-4:t-1}\rvert$	0.183***	(0.032)
$\overline{IndReturn}_{i,t-4:t-1}$	0.100	(0.071)	$\lvert\overline{IndReturn}_{i,t-4:t-1}\rvert$	−0.037	(0.041)
$\overline{MktCap}_{i,t-4:t-1}$	−0.862***	(0.197)	$\overline{MktCap}_{i,t-4:t-1}$	0.465***	(0.080)
$\overline{Volume}_{i,t-4:t-1}$	0.297	(0.154)	$\overline{Volume}_{i,t-4:t-1}$	−0.866***	(0.104)
$\overline{P2B}_{i,t-4:t-1}$	−2.077	(1.033)	$\lvert\overline{P2B}_{i,t-4:t-1}\rvert$	2.098***	(0.519)
$\overline{Turnover}_{i,t-4:t-1}$	−1.038**	(0.351)	$\overline{Turnover}_{i,t-4:t-1}$	0.914***	(0.140)
Constant	30.700***	(7.245)	Constant	−7.317	(3.964)
Observations	90,864		Observations	90,864	
Number of weeks	41		Number of weeks	41	

Note. Newey-West adjusted standard errors are in parentheses.
$^{**}p < 0.01$; $^{***}p < 0.001$.

Table 5. Trading Strategy: Results of Two-Way Sorts and Multivariate Regressions

Panel A. Tercile trading profits						
EAC	Mkt-Rf	SMB	HML	RMW	CMA	Alpha
T1 (low)	−0.182*	−0.492***	−0.450***	0.430*	0.819***	0.0024*
	(0.086)	(0.140)	(0.119)	(0.203)	(0.222)	(0.0011)
T2	−0.034	−0.482***	−0.179*	0.173	0.406**	0.0022**
	(0.057)	(0.106)	(0.076)	(0.149)	(0.126)	(0.0008)
T3 (high)	−0.080	−0.037	0.278	0.225	−0.079	−0.0013
	(0.098)	(0.100)	(0.150)	(0.136)	(0.212)	(0.0011)

Panel B. Quintile trading profits						
EAC	Mkt-Rf	SMB	HML	RMW	CMA	Alpha
Q1 (low)	−0.321***	−0.551***	−0.577***	0.545**	1.145***	0.0033***
	(0.055)	(0.109)	(0.095)	(0.198)	(0.159)	(0.0007)
Q2	−0.231**	−0.722***	−0.343*	−0.0504	0.666*	0.0016
	(0.072)	(0.148)	(0.154)	(0.209)	(0.264)	(0.0010)
Q3	0.022	−0.416*	−0.158	0.485	0.755*	0.0012
	(0.124)	(0.162)	(0.203)	(0.252)	(0.313)	(0.0015)
Q4	−0.072	0.111	0.142	0.440***	0.079	−0.0005
	(0.073)	(0.108)	(0.099)	(0.125)	(0.140)	(0.0009)
Q5 (high)	−0.049	0.018	0.309***	0.111	−0.049	−0.0020***
	(0.032)	(0.068)	(0.079)	(0.085)	(0.115)	(0.0005)

Note. Standard errors are in parentheses.
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

rebalanced every week using a sliding window approach. The resulting weekly portfolio returns are then regressed on Fama-French's risk factors using ARCH (i.e., autoregressive conditional heteroscedasticity) estimation.

The benchmark of our trading strategy is against the market expectation. If portfolio returns can be fully explained by the market risk factors, the estimated intercept of the model, that is, the alpha, would be insignificant. Instead, if the portfolios based on our trading strategy can yield significantly higher returns over the market expectation, the model would produce a significantly positive alpha.

From Table 5, we find that portfolios under low-EAC scenarios (i.e., T1, T2, and Q1) have a significant and positive alpha. In a low-EAC condition (i.e., T1 and Q1), our trading strategy can yield a raw weekly portfolio return of 24 and 33 basis points. This provides evidence on the existence of excess returns that cannot be explained by the widely accepted market risk factors but by our EAC measure. This result suggests that a portfolio derived from investors' buying positive sentiment stocks and selling negative sentiment stocks can earn significant excess returns under the circumstance of low EAC. We also find that under quintile 5 (Q5), the portfolio yields negative returns. This further supports our findings that high EAC is associated with a large magnitude of return reversal because the fast information diffusion facilitates the incorporation of information into stock prices, which leads to a sharp return reversal.

6. Predictive Performance of Different Information Sources

We formulate EAC based on composite information from various channels. It is possible that different information has different levels of importance on return prediction. To verify this, we first investigate whether EAC based on an individual information source can predict stock returns. Then, we study information from social and traditional media, respectively, and compare their relative importance on return prediction. Finally, we compare the predictive performance of our EAC metric with that of an attention measure to determine the relative importance of attention and coattention on predicting abnormal returns.

6.1. EACs Based on Separate Information Sources

To compare the relative importance of different information sources, we formulate a set of EACs using individual attention proxies (i.e., number of reads, number of comments, news coverage, and analyst reports) respectively instead of the composite attention index. Similarly, to allow for out-of-sample validation, we partitioned our data set into a training set that contained the prior 40 weeks and a test set that included the remaining 14 weeks for performance evaluation. Table 6 shows the predictive results for EACs based on different information sources.

From Table 6, we find that in most cases, EAC based on composite information outperforms EACs based on separate information sources in predicting both the direction and magnitude of abnormal returns. In addition, EAC based on investor reads and comments tend to generate higher predictive accuracy than EAC based on news and reports. This suggests that information from social media channels can yield better predictions compared with information from traditional media. Interestingly, under some prediction methods, EAC based on investor reads is slightly better than the composite EAC in predicting abnormal returns. As the number of reads of a focal stock refers to the volume of investor reads of stock-related posts on the online stock forum, it captures the intensity of social media information absorbed and processed by investors.[17] Therefore, we conjecture that the predictability of EAC on stock returns is dominated by information from social media instead of traditional media such as news and reports.

6.2. Social Media vs. Traditional Media

To formally show the relative importance of different information channels, we separate the multiple information sources into social media information (e.g., investor reads and comments) and traditional media information (e.g., analyst reports, financial announcements), and

Table 6. Predictive Accuracy of EACs Based on Different Information Sources

Panel A. Fama-MacBeth regressions with Newey-West adjusted standard errors

Models	Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it} \rvert$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
EAC (*Composite*)	4.8733	3.4669	2.9213	3.8953	3.2360	10.9820
EAC (*Read*)	4.8442	3.4301	2.9539	3.9004	3.2479	13.6322
EAC (*Comment*)	5.0037	3.6577	3.9176	3.9944	3.3693	11.7604
EAC (*News*)	5.0388	3.7219	4.4631	4.0699	3.4572	12.7647
EAC (*Report*)	5.0962	3.7994	4.9319	4.0968	3.4893	12.9266

Panel B. Support vector regressions with polynomial kernels

Models	Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it} \rvert$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
EAC (*Composite*)	4.6994	3.2276	1.7192	3.2719	2.2410	6.7032
EAC (*Read*)	4.7625	3.2807	1.7204	3.2956	2.2832	6.9549
EAC (*Comment*)	4.7563	3.3209	1.9740	3.2956	2.2831	6.9540
EAC (*News*)	4.7857	3.2868	1.9632	3.2984	2.3011	7.1715
EAC (*Report*)	4.7660	3.3534	2.1954	3.3028	2.3213	7.3821

Panel C. Neural networks using multilayer perceptron

Models	Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it} \rvert$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
EAC (*Composite*)	4.8984	3.4601	2.4535	3.3775	2.4939	8.7281
EAC (*Read*)	4.9316	3.5627	4.1182	3.3674	2.4439	8.1883
EAC (*Comment*)	5.3402	3.7318	4.1488	3.3710	2.5370	8.8576
EAC (*News*)	5.3782	3.9373	5.5352	3.5068	2.7530	9.8475
EAC (*Report*)	5.6085	3.9317	4.9789	3.5873	2.8022	9.8968

Panel D. Decision trees

Models	Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it} \rvert$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
EAC (*Composite*)	4.6728	3.2464	1.1345	3.6925	3.0288	11.3632
EAC (*Read*)	4.7619	3.3114	1.0500	3.7239	3.0281	11.2199
EAC (*Comment*)	4.7586	3.3140	1.3769	3.7681	3.1208	11.7462
EAC (*News*)	4.8114	3.3602	1.5880	4.1842	3.5589	13.5167
EAC (*Report*)	4.7906	3.3356	1.9842	4.0923	3.4891	13.2546

Panel E. Random forests

Models	Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it} \rvert$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
EAC (*Composite*)	4.6871	3.2362	1.3078	3.7007	2.9988	11.0889
EAC (*Read*)	4.7439	3.2974	1.3377	3.8750	3.2251	12.0951
EAC (*Comment*)	4.7537	3.2969	1.3718	3.8790	3.2275	12.1069
EAC (*News*)	4.7215	3.3098	1.6575	3.9878	3.3751	12.8493
EAC (*Report*)	4.7214	3.3123	1.6455	3.9897	3.3782	12.8628

Panel F. Gradient-boosted decision trees

Models	Model (1): $AbnReturn_{it}$			Model (2): $\lvert AbnReturn_{it} \rvert$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
EAC (*Composite*)	4.6762	3.2408	1.1675	3.6859	2.9936	10.9883
EAC (*Read*)	4.7427	3.3009	1.1820	3.8179	3.1776	11.9079
EAC (*Comment*)	4.7428	3.3004	1.1756	3.8213	3.1812	11.9097
EAC (*News*)	4.7616	3.3045	1.2298	4.0049	3.4129	12.9788
EAC (*Report*)	4.7543	3.3008	1.2036	4.0336	3.4445	13.1132

evaluate the predictive performance of EAC based on the two media types, respectively. Please refer to Online Appendix G for details on the construction of EAC (*Social Media*) and EAC (*Traditional Media*). Similarly, we assess the out-of-sample predictive accuracy and find that EAC (*Composite*) and EAC (*Social Media*) yield more accurate predictions than EAC (*Traditional Media*) (see Table G2 of Online Appendix G for results). To further investigate the extent to which EACs based on composite and social media information outperform EAC based on traditional media information, we adopt the bootstrapping approach in Section 5.2 to calculate the 95% confidence intervals for the difference in mean absolute errors between various models. Table 7 shows the results.

From panel A of Table 7, we find that all the confidence intervals are greater than zero, which suggests that EAC (*Composite*) can significantly outperform EAC (*Traditional Media*) in predicting abnormal returns. Besides, as shown in panel B, EAC (*Social Media*) tends to have better predictive performance than EAC (*Traditional Media*). The 95% confidence intervals in panel C includes zero in most cases, which further indicates that the predictive power of EAC on abnormal returns is primarily from social media instead of traditional media. We also compare the predictive performance of EAC based on social media with that of an attention measure based on social media information. Panel D of Table 7 shows the corresponding results. We find that EAC (*Social Media*) can significantly outperform attention (*Social Media*) in predicting abnormal returns.[18] This suggests that coattention from other stocks plays an important role above and beyond direct attention in predicting abnormal returns. Thus, our analysis expands on previous studies (e.g., Luo et al. 2013, Yu et al. 2013, Chen et al. 2014) that focus on attention-based measures of social media for stock performance prediction.

Table 7. Comparisons of EACs based on Social and Traditional Media

Panel A. 95% CIs for MAE[*EAC (Traditional Media)*] − MAE[*EAC (Composite)*]				
Method	(1) Predicting $AbnReturn_{it}$		(2) Predicting $\|AbnReturn_{it}\|$	
Fama-MacBeth	0.2231	0.3325	0.0303	0.0970
SVR	0.0700	0.1832	0.0038	0.0785
MLP	0.2926	0.4239	0.3072	0.3834
DT	0.0357	0.1494	0.4684	0.5370
RF	0.0076	0.1205	0.1919	0.2639
GBDT	0.0100	0.1192	0.3188	0.3900
Panel B. 95% CIs for MAE[*EAC* (*Traditional Media*)] − MAE[*EAC* (*Social Media*)]				
Method	(1) Predicting $AbnReturn_{it}$		(2) Predicting $\|AbnReturn_{it}\|$	
Fama-MacBeth	0.2174	0.3145	0.0302	0.0917
SVR	0.0166	0.1260	−0.0274	0.0528
MLP	0.0252	0.1526	0.3309	0.4036
DT	−0.0221	0.0863	0.4997	0.5697
RF	0.0078	0.1110	0.1852	0.2528
GBDT	−0.0553	0.0577	0.3183	0.3844
Panel C. 95% CIs for MAE[*EAC (Social Media)*] − MAE[*EAC (Composite)*]				
Method	(1) Predicting $AbnReturn_{it}$		(2) Predicting $\|AbnReturn_{it}\|$	
Fama-MacBeth	−0.0454	0.0607	−0.0304	0.0348
SVR	−0.0063	0.1108	−0.0153	0.0644
MLP	0.2090	0.3204	−0.0581	0.0154
DT	−0.0066	0.1106	−0.0663	0.0048
RF	−0.0455	0.0704	−0.0293	0.0432
GBDT	0.0038	0.1160	−0.0352	0.0394
Panel D. 95% CIs for MAE[*Attention* (*Social Media*)] − MAE[*EAC* (*Social Media*)]				
Method	(1) Predicting $AbnReturn_{it}$		(2) Predicting $\|AbnReturn_{it}\|$	
Fama-MacBeth	0.1936	0.2961	0.2529	0.3144
SVR	0.1138	0.2184	0.0420	0.1170
MLP	0.0993	0.2089	0.1321	0.2026
DT	0.1367	0.2467	0.1520	0.2176
RF	0.1723	0.2791	0.1235	0.1898
GBDT	0.1192	0.2187	0.0322	0.0949

7. Discussion and Conclusion

This research, to our knowledge, is the first attempt to capture and model the information flows of different sources among a network of assets derived from visitors' digital footprints and to study their market impact for individual assets. We do so by developing a novel composite measure, EAC, which captures direct information to nodes (i.e., attention) and information flows due to coattention across nodes in a network. We validate the effectiveness of this measure on predicting outcomes by its application in the stock market. The EAC accounts for both attention to individual stocks (e.g., investor reading of stock-related posts, comments posted, news articles, and analyst reports) and coattention across networks of connected stocks based on investors' coattention obtained from their digital footprints.

Our study is closely aligned with the design science paradigm under which the design artifact (i.e., EAC) should be rigorously evaluated to demonstrate its usefulness by providing solutions and understanding to a problem domain (Hevner et al. 2004, Gregor and Hevner 2013). To achieve these goals, we show that EAC is a powerful predictor for stock performance in which we focus on its predictability on both the direction and magnitude of abnormal returns. We develop and follow a multifacet validation strategy to assess the predictive performance of EAC. Based on a set of linear and nonlinear methods, we first compare models with and without EAC and highlight that EAC can provide additional predictive gains to stock returns. We then compare EAC with a set of alternative metrics such as degree, closeness, betweenness, and PageRank centralities and demonstrate its superior performance in return prediction. We also show that our EAC specification is better than alternative specifications and our results are robust and consistent under different matrix normalization methods and censored data situations. We further evaluate the effectiveness of EAC by proposing a trading strategy. We find that in low-EAC scenarios, investors can gain profits by holding a portfolio through buying positive sentiment stocks and selling negative sentiment stocks. However, the likelihood to gain profits in high-EAC conditions is minimal due to the price efficiency. This provides support to the gradual information diffusion (Hong and Stein 1999) and investor recognition hypothesis (Merton 1987) under which the low-EAC condition generates excess portfolio returns. These findings are relevant to market practitioners and can trigger theoretical thinking about the role of information flows in altering return predictability and price efficiency. In addition, our results reveal that the combined information flows from multiple sources yield better predictions compared with individual information source, and social media has greater predictive power on abnormal returns than traditional media. This is possible because social media tends to have higher visibility and information richness and is more "socially contagious" than traditional media (Luo et al. 2013, Yu et al. 2013). We also establish the relative importance of coattention and attention in predicting abnormal returns and uncover that coattention from other stocks can provide additional predictive gains beyond direct attention.

We provide important implications for practitioners. First, the findings of our trading strategy can provide guidance on portfolio management for investors in many ways. For example, an EAC index for stocks can be constructed to capture the trace of information diffusion in the stock market. The time series of that EAC index may also provide insights into the dynamics of information diffusion in the stock market. Second, this study offers practical implications on strategic actions in developing new investment portfolios based on information flows using EAC. What could be very interesting is that if the real clickstream data could be used and stock networks can be created at the granularity of millisecond or even faster, our findings may help generate new high-frequency trading algorithms. Third, our work also shows the opportunity for infomediaries to provide more value-added information in different formats/types and from different sources to market participants.

Our EAC metric can be generalized to other contexts as well. Consider a coview network where nodes are different products. Products in the network have different levels of popularities as reflected by the search volume, user ratings, number of reviews, and so on. The edges in the network are also heterogeneous because some products are coviewed more intensively or copurchased more frequently than others, so there would be more information flows between these products. Then, an EAC metric can be constructed by considering both the node and edge heterogeneity to reflect the level of the overall popularity of individual products in the network, and the EAC metric in the product network can be used to forecast consumer demand or product sales. We show additional examples in Online Appendix H.

Although this research provides several notable findings and contributions, there are some limitations and opportunities that could be addressed in future work. First, there could be other metrics of coattention such as actual coattention frequencies, investors' trading habits (e.g., how investors buy stocks) and reading and commenting on a set of stocks. Future research may investigate the impact of alternative measures of coattention on information flows and stock returns. Besides, this work primarily focuses on the volume of information from different sources. One could incorporate word-of-mouth expressed in comments from user-generated content within the information flows.

Acknowledgments

The authors thank the senior editor, associate editor, and three anonymous reviewers for their invaluable and insightful comments that have greatly helped to improve this work.

Endnotes

[1] Henceforth, we use only assets to represent both products and assets.

[2] We focus on all the A-share stocks listed in both Shanghai and Shenzhen Stock Exchanges that are denominated in RMB (Chinese currency) and traded by Mainland Chinese citizens. The Chinese stock market is heavily dominated by individual and retail investors who account for over 80% of transactions on stock exchanges (Shen and Takada 2016).

[3] The four information sources include: (1) the number of times messages for a stock are read by investors, (2) the number of investor comments related to a stock, (3) the number of news articles about a stock, and (4) the number of analyst reports on a stock.

[4] Prior works (e.g., Dorogovtsev and Mendes 2003, Carletti and Righi 2010, Dai et al. 2016) adopt similar network settings by assigning weights to nodes to reflect the importance or popularity of individuals in a network and assigning edge weights to represent the intensity of links or relationships among nodes.

[5] Because both network structure and information environment can change over time, analysis of a series of networks over different time periods allows us to capture these dynamic changes.

[6] Da et al. (2011) use search volume index as a measure of attention. However, equivalent data for the Chinese stock market were not available at the time we collected the data.

[7] The first principal factor has the largest eigenvalue. The scree plot of eigenvalues indicates that an inflection point occurs after the first factor.

[8] Calculating the weighted average (especially value-weighted average) of a proxy is common in finance studies (e.g., Grinblatt and Titman 1989, Grinblatt et al. 1995, Maillard et al. 2010), which accounts for the time-varying dependency on the market capitalization of a focal stock.

[9] Although R_{ij}^2 should equal R_{ji}^2 based on the calculations using Equation (1), the coattention network is not symmetric because the two stocks *i* and *j* may not be in each other's coattention list simultaneously according to the platform's data-generation scheme as described in endnote 10. When there is no observed coattention, there is no edge and the corresponding edge weight is set to zero.

[10] Sina Finance provides the following information about the data-generation process: Users of Sina Finance can follow the stocks of interest and add them into their "favorites" list. Based on the "favorites" lists of all the users, Sina determines the coattention list by identifying another nine stocks that have the highest coattention frequencies with the focal stock, and these nine stocks constitute the coattention list shown on the homepage of the focal stock.

[11] Sina does not use recommendation schemes in generating coattention lists.

[12] The site eastmoney.com is one of the most popular Chinese financial portals. It has an online forum of listed firms.

[13] One benefit of the bootstrapping method is that we do not need to assume the population has a normal distribution. We also calculate confidence intervals for mean differences in MAPE between different models, and the results are shown in Online Appendix F.4.

[14] There is evidence from prior research that positive (negative) sentiment is associated with positive (negative) returns (Tetlock et al. 2008).

[15] We also report the results of the trading strategy using the in-sample in Online Appendix F.6. The results are consistent with those of the out-of-sample trading strategy in Table 5.

[16] T1 to T3 correspond to low-EAC, medium-EAC, and high-EAC stocks, respectively. Q1 to Q5 correspond to lowest-, low-, medium-, high-, highest-EAC stocks, respectively.

[17] Although investor reads and comments are both from social media channels, we find that EAC(*Read*) tends to perform better than EAC(*Comment*). A possible explanation is that compared with reading a post, making a comment requires more time and behavioral costs (e.g., register an account and log on to the account). Thus, the intensity of comment is much lower and varies less than the intensity of investor read, which results in better predictive performance of EAC based on investor read.

[18] We also compare the predictive performance of the composite EAC metric and attention measures based on composite, social media, and traditional media information. The results are shown in Table G3, G4 of Online Appendix G, which suggest that the composite EAC metric also outperforms composite attention measure and attention based on social and traditional media information.

References

Agarwal A, Leung ACM, Konana P, Kumar A (2017) Cosearch attention and stock return predictability in supply chains. *Inform. Systems Res.* 28(2):265–288.

Antweiler W, Frank MZ (2004) Is all that talk just noise? The information content of internet stock message boards. *J. Finance* 59(3):1259–1294.

Baker M, Wurgler J (2007) Investor sentiment in the stock market. *J. Econom. Perspect.* 21(2):129–151.

Barber BM, Odean T (2008) All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Rev. Financial Stud.* 21(2):785–818.

Belo F, Lin X, Bazdresch S (2014) Labor hiring, investment, and stock return predictability in the cross section. *J. Political Econom.* 122(1):129–177.

Borgatti SP (2005) Centrality and network flow. *Soc. Networks* 27(1): 55–71.

Brown GW, Cliff MT (2004) Investor sentiment and the near-term stock market. *J. Empirical Finance* 11(1):1–27.

Carletti T, Righi S (2010) Weighted fractal networks. *Physica A* 389(10):2134–2142.

Chen H, De P, Hu YJ, Hwang B-H (2014) Wisdom of crowds: The value of stock opinions transmitted through social media. *Rev. Financial Stud.* 27(5):1367–1403.

Da Z, Engelberg J, Gao P (2011) In search of attention. *J. Finance* 66(5):1461–1499.

Dai M, Ye D, Hou J, Xi L, Su W (2016) Average weighted trapping time of the node- and edge- weighted fractal networks. *Comm. Nonlinear Sci. Numer. Simulation* 39:209–219.

Dewan S, Ren F (2007) Risk and return of information technology initiatives: Evidence from electronic commerce announcements. *Inform. Systems Res.* 18(4):370–394.

Dhar V, Geva T, Oestreicher-Singer G, Sundararajan A (2014) Prediction in economic networks. *Inform. Systems Res.* 25(2): 264–284.

Dietterich TG (2000) Ensemble methods in machine learning. *Internat. Workshop Multiple Classifier Systems* (MCS 2000) (Springer, Berlin, Heidelberg), 1–15.

Dorogovtsev SN, Mendes JFF (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford).

Drake MS, Jennings J, Roulstone DT, Thornock JR (2016) The comovement of investor attention. *Management Sci.* 63(9): 2847–2867.

Fama EF, MacBeth JD (1973) Risk, return, and equilibrium: Empirical tests. *J. Political Econom.* 81(3):607–636.

Gregor S, Hevner AR (2013) Positioning and presenting design science research for maximum impact. *MIS Quart.* 37(2): 337–355.

Grinblatt M, Titman S (1989) Mutual fund performance: An analysis of quarterly portfolio holdings. *J. Bus.* 62(3):393–416.

Grinblatt M, Titman S, Wermers R (1995) Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior. *Amer. Econom. Rev.* 85(5): 1088–1105.

Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. *MIS Quart.* 28(1):75–105.

Hirshleifer D (2015) Behavioral finance. *Annual Rev. Financial Econom.* 7:133–159.

Hirshleifer D, Teoh SH (2003) Limited attention, information disclosure, and financial reporting. *J. Accounting Econom.* 36(1-3):337–386.

Hong H, Stein JC (1999) A unified theory of underreaction, momentum trading, and overreaction in asset markets. *J. Finance* 54(6): 2143–2184.

Kuncheva LI (2014) *Combining Pattern Classifiers: Methods and Algorithms*, 2nd ed. (Wiley, Hoboken, NJ).

Leung ACM, Agarwal A, Konana P, Kumar A (2017) Network analysis of search dynamics: The case of stock habitats. *Management Sci.* 63(8):2667–2687.

Lin Z, Goh KY, Heng C-S (2017) The demand effects of product recommendation networks: An empirical analysis of network diversity and stability. *MIS Quart.* 41(2):397–426.

Liu Y, Pant G, Sheng ORL (2020) Predicting labor market competition: Leveraging interfirm network and employee skills. *Inform. Systems Res.* 31(4):1443–1466.

Luo X, Zhang J, Duan W (2013) Social media and firm equity value. *Inform. Systems Res.* 24(1):146–163.

Maillard S, Roncalli T, Teïletche J (2010) The properties of equally weighted risk contribution portfolios. *J. Portfolio Management* 36(4):60–70.

Menzly L, Ozbas O (2010) Market segmentation and cross-predictability of returns. *J. Finance* 65(4):1555–1580.

Merton RC (1987) A simple model of capital market equilibrium with incomplete information. *J. Finance* 42(3):483–510.

Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55(3):703–708.

Newman ME (2010) Mathematics of networks. chapter 6. 1st ed. *Networks: An Introduction* (Oxford University Press, Oxford, England), 109–157.

Oestreicher-Singer G, Sundararajan A (2012) Recommendation networks and the long tail of electronic commerce. *MIS Quart.* 36(1):65–83.

Oestreicher-Singer G, Libai B, Sivan L, Carmi E, Yassin O (2013) The network value of products. *J. Marketing* 77(3):1–14.

Peng G, Dey D (2013) Research note—A dynamic view of the impact of network structure on technology adoption: The case of OSS development. *Inform. Systems Res.* 24(4):1087–1099.

Pontiff J, Woodgate A (2008) Share issuance and cross-sectional returns. *J. Finance* 63(2):921–945.

Rokach L (2010) Ensemble-based classifiers. *Artificial Intelligence Rev.* 33(1):1–39.

Sabherwal S, Sarkar SK, Zhang Y (2011) Do internet stock message boards influence trading? Evidence from heavily discussed stocks with no fundamental news. *J. Bus. Finance Accounting* 38(9-10):1209–1237.

Shen S, Takada K (2016) Giving up on stocks: China's retail investors seek safety first. Reuters (January 14), https://www.reuters.com/article/us-china-markets-retail/giving-up-on-stocks-chinas-retail-investors-seek-safety-first-idUSKCN0US0JS20160114.

Susarla A, Oh J-H, Tan Y (2012) Social networks and the diffusion of user-generated content: Evidence from YouTube. *Inform. Systems Res.* 23(1):23–41.

Tetlock PC, Saar-Tsechansky M, Macskassy S (2008) More than words: Quantifying language to measure firms' fundamentals. *J. Finance* 63(3):1437–1467.

Wahal S, Yavuz MD (2013) Style investing, comovement and return predictability. *J. Financial Econom.* 107(1):136–154.

Yu Y, Duan W, Cao Q (2013) The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems* 55(4):919–926.

Zhang X, Wang C (2012) Network positions and contributions to online public goods: The case of Chinese Wikipedia. *J. Management Inform. Systems* 29(2):11–40.

Zhang X, Zhang L (2015) How does the Internet affect the financial market? An equilibrium model of Internet-facilitated feedback trading. *MIS Quart.* 39(1):17–38.